A Rigorous Framework for Validating Ensemble Forecasts

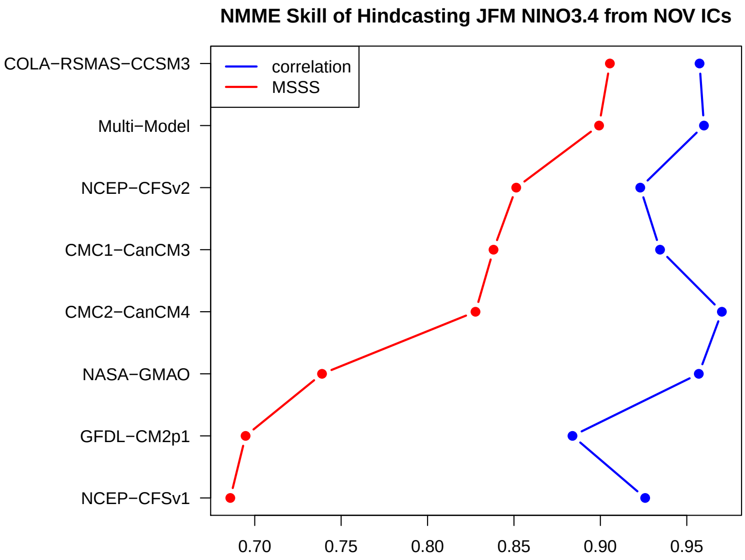
Timothy DelSole

George Mason University, Fairfax, Va and
Center for Ocean-Land-Atmosphere Studies, Calverton, MD

July 30, 2013

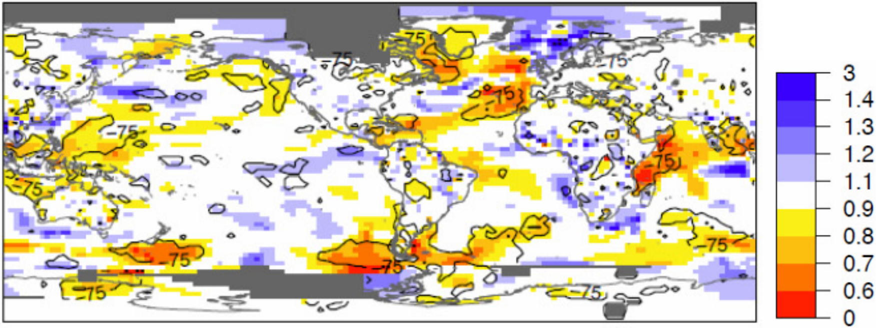
Collaborators: Michael K. Tippett (IRI) and Jyothi Nattalla (COLA/GMU)

How Do You Compare Forecast Skill/Scores?



Compare Mean Square Error?

$$\frac{RMSE_{Init}}{RMSE_{NoInit}}$$



From fig. 11.7 of AR5; ratio of rmse between initialized and non-initialized decadal hindcasts for years 2-5. Dots show 5% significant difference based on one-sided F-test.

Test Equality of Variance ($\sigma_1^2 = \sigma_2^2$)

Statistic: Let s_1^2 and s_2^2 be the sample variances:

$$F = \frac{s_1^2}{s_2^2}.$$

Theorem: If samples are independent and identically distributed as a Gaussian, then

$$F \sim F_{\nu_1,\nu_2}.$$

where ν_1 and ν_2 are the appropriate degrees of freedom.

Test Equality of Variance ($\sigma_1^2 = \sigma_2^2$)

Statistic: Let s_1^2 and s_2^2 be the sample variances:

$$F = \frac{s_1^2}{s_2^2}.$$

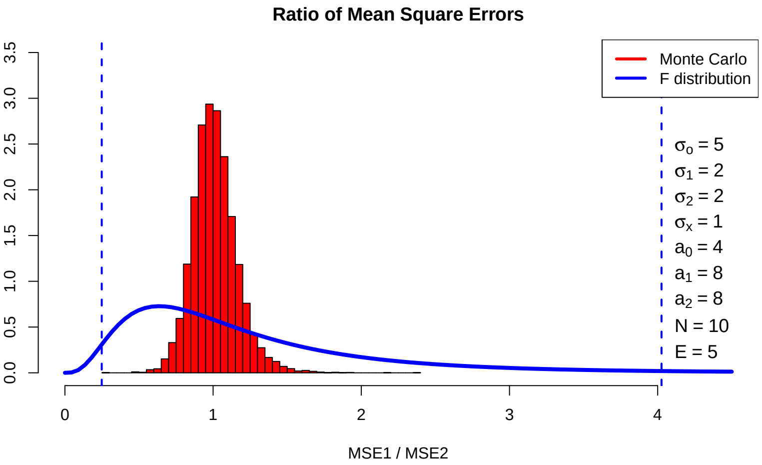
Theorem: If samples are **independent** and identically distributed as a Gaussian, then

$$F \sim F_{\nu_1,\nu_2}.$$

where ν_1 and ν_2 are the appropriate degrees of freedom.

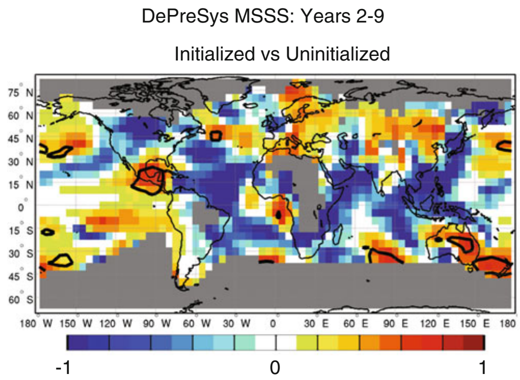
Idealized Forecast/Observation System

$$\text{observation} = a_o \text{ signal} + \text{noise}_o$$
$$\text{forecast } 1 = a_1 \text{ signal} + \text{noise}_1$$
$$\text{forecast } 2 = a_2 \text{ signal} + \text{noise}_2$$



Compare Mean Square Error?

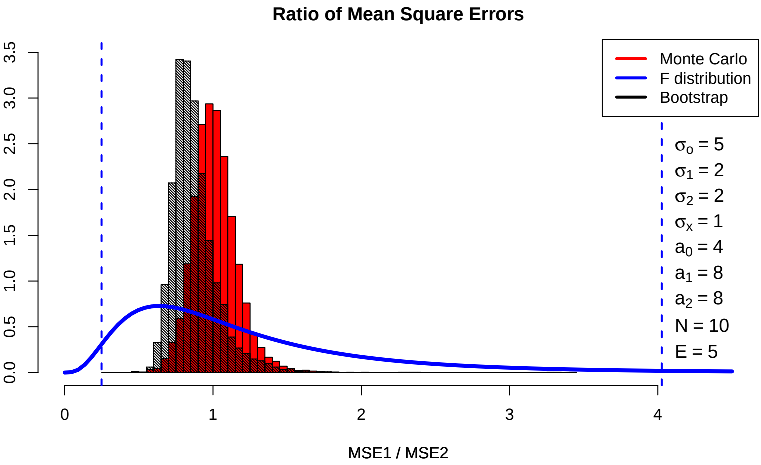
$$MSSS = 1 - \frac{RMSE_{Init}}{RMSE_{NoInit}}$$



from Goddard et al. (2012). Contour line "indicates significance that MSSS is positive at 95% confidence level," based on bootstrap method.

Idealized Forecast/Observation System

$$\text{observation} = a_o \ \text{signal} + \text{noise}_o$$
$$\text{forecast } 1 = a_1 \ \text{signal} + \text{noise}_1$$
$$\text{forecast } 2 = a_2 \ \text{signal} + \text{noise}_2$$



Conclusion

A rigorous significance test of skill differences does not exist when

- validation measure is calculated using the same verification.
- the prediction models are not nested.

Rigorous Ways to Compare Forecasts

Compare nested prediction models.

or

Confirm that observations are consistent with forecast distribution, then test differences in forecast spread.

or

Compare skills estimated from independent verifications.

Does the Multi-Model Ensemble Enhance Skill?

Consider the nested regression models

single model	O	$=$	a	F_i	$+$			$+$	ϵ
combination	O	$=$	a	F_i	$+$	b	M_i	$+$	ϵ
	obs			forecast model i			multimodel mean except i		error

Does the Multi-Model Ensemble Enhance Skill?

Consider the nested regression models

single model	O	$=$	a	F_i	$+$		$+$	ϵ
combination	O	$=$	a	F_i	$+$	$b \quad M_i$	$+$	ϵ
	obs			forecast model i		multimodel mean except i		error

Is MSE[combination] $<$ MSE[single model] ?

We are assessing whether **combining** two forecasts significantly improves the forecast **relative to one forecast**.

We are **not** assessing whether one forecast **in isolation** is significantly better than another.

Equivalence

$$\begin{array}{lccccccccc} \text{single model} & O & = & a & F_i & + & & & + & \epsilon \\ \text{combination} & O & = & a & F_i & + & b & M_i & + & \epsilon \end{array}$$

Testing the hypothesis

$$\mathbf{MSE[combination] = MSE[single\ model]}$$

is equivalent to testing the hypothesis

$$\mathbf{b = 0}$$

Hypothesis Test

If the null hypothesis $b = 0$ is true, then

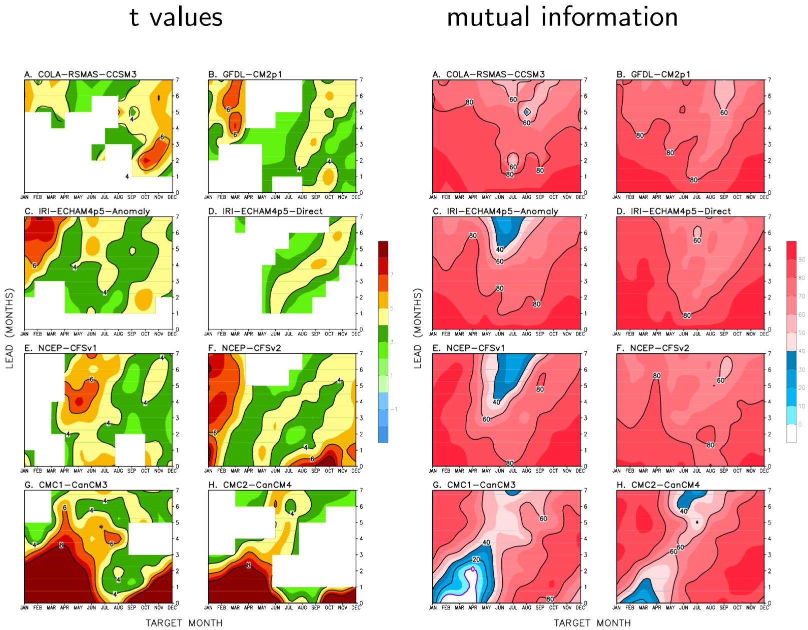
$$t = \frac{b_{\text{least squares}}}{\sigma_b}$$

has a t distribution with $N - 3$ degrees of freedom.

National Multi-Model Ensemble

- Hindcasts initialized every month from 1982-2010
- At least 6 month lead
- Analyze NINO3.4
- Separate climatologies for 1982-1999 and 2000-2010
- Verification: OISST

model	ensemble size
CMC1-CanCM3	10
CMC2-CanCM4	10
COLA-RSMAS-CCSM3	6
GFDL-CM2p1	10
NASA-GMAO	11
NCEP-CFSv1	15
NCEP-CFSv2	24


t values
mutual information
A. COLA–RSMAS–CCSM3
B. GFDL–CM2p1
C. IRI–ECHAM4p5–Anomaly
D. IRI–ECHAM4p5–Direct
E. NCEP–CFSv1
F. NCEP–CFSv2
G. CMC1–CanCM3
H. CMC2–CanCM4
LEAD (MONTHS)
TARGET MONTH
JAN FEB MAR APR MAY JUN JUL AUG SEP OCT NOV DEC

Second Approach: Compare Calibrated Forecasts

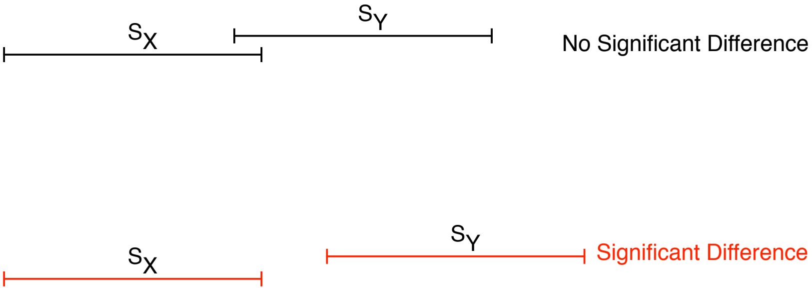
If observations are drawn from the forecast distribution, then

$$\frac{1}{1 + \frac{1}{E}} \langle MSE \rangle = \sigma^2_{forecast}$$

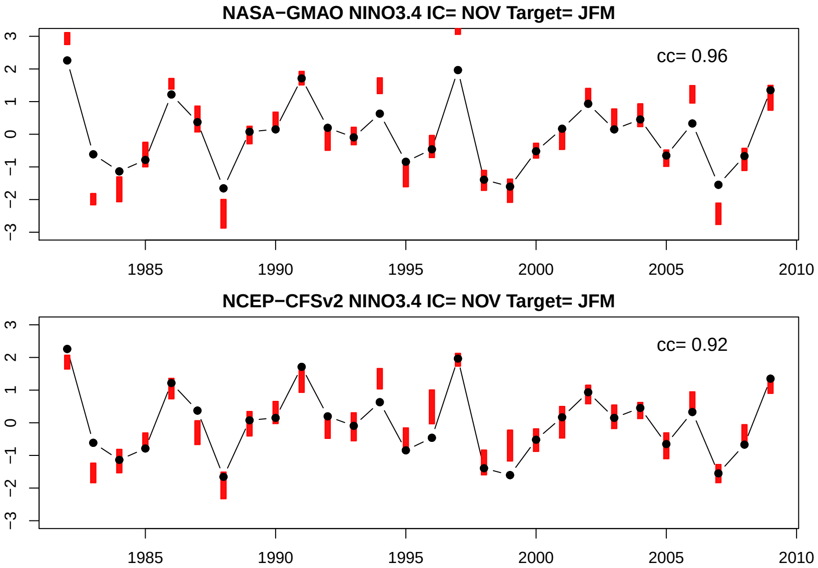
If the calibration hypothesis cannot be rejected, then a significantly better forecast would have significantly smaller noise:

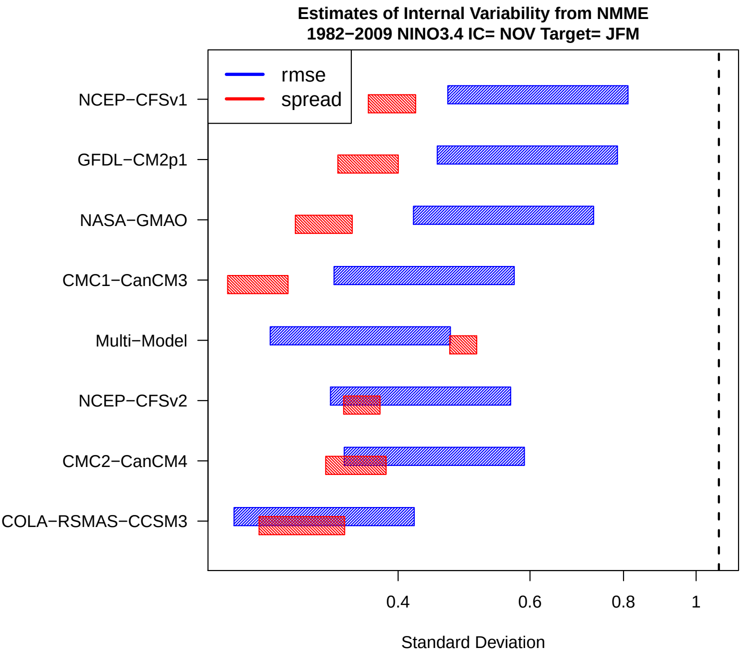
$$\sigma^2_{forecast,1} < \sigma^2_{forecast,2}$$

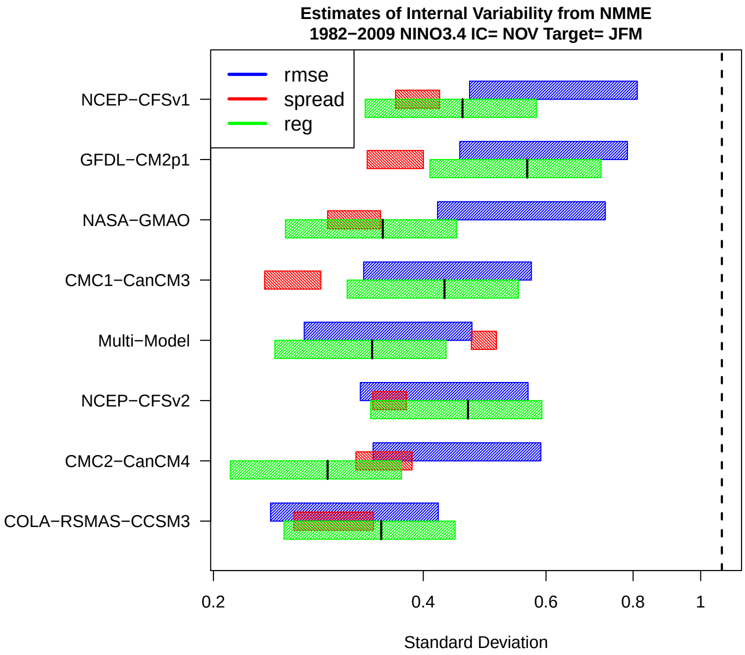
Confidence Interval for Variance



The standard 95% confidence interval for variance can be modified slightly to correspond precisely to an F-test for equality of variance.


NASA−GMAO NINO3.4 IC= NOV Target= JFM
cc= 0.96
NCEP−CFSv2 NINO3.4 IC= NOV Target= JFM
cc= 0.92
3
2
1
0
−1
−2
−3
1985
1990
1995
2000
2005
2010


Estimates of Internal Variability from NMME
1982−2009 NINO3.4 IC= NOV Target= JFM
rmse
spread
NCEP−CFSv1
GFDL−CM2p1
NASA−GMAO
CMC1−CanCM3
Multi−Model
NCEP−CFSv2
CMC2−CanCM4
COLA−RSMAS−CCSM3
0.4
0.6
0.8
1
Standard Deviation



Estimates of Internal Variability from NMME
1982−2009 NINO3.4 IC= NOV Target= JFM
rmse
spread
reg
NCEP−CFSv1
GFDL−CM2p1
NASA−GMAO
CMC1−CanCM3
Multi−Model
NCEP−CFSv2
CMC2−CanCM4
COLA−RSMAS−CCSM3
0.2
0.4
0.6
0.8
1
Standard Deviation

Summary

- Testing the significance of a difference in skill is difficult because
  - skills are not independent
  - dynamical prediction models are not nested
- The bootstrap distribution is sensitive to the sample that actually occurs, even for large bootstrap samples.
- Two ways to rigorously compare skills:
  - compare skills calculated from independent verifications
  - test calibration, then test differences in forecast spread
- For National Multi-Model Ensemble
  - MSE intervals are large and overlap with each other
  - MSE consistent with forecast spread for 4 models.
  - Of these, multi-model forecast has significantly worse score.
- Proposed method for deciding whether multi-model enhances skill.
  - Every model has periods in which multi-model enhances skill.
  - Multi-model systematically improves skill during spring barrier.